HUMAN FACTOR PROBLEMS IN ANTI-SUBMARINE WARFARE

Technical Report 4

AN EXPLORATORY STUDY OF THE CORRELATES OF VIGILANCE PERFORMANCE

HUMAN FACTORS RESEARCH, INCORPORATED
1112 Crenshaw Boulevard • Los Angeles 19, California • WEbster 3-7356

HUMAN FACTOR PROBLEMS IN ANTI-SUBMARINE WARFARE

Technical Report 4

AN EXPLORATORY STUDY OF THE CORRELATES OF VIGILANCE PERFORMANCE

James J. McGrath
Albert Harabedian
Donald N. Buckner

Prepared for

Personnel and Training Branch
Psychological Sciences Division
Office of Naval Research
Department of the Navy

by

Human Factors Research, Incorporated
Los Angeles 19, California

February 1960
Contract Nonr 2649(00)
NR 153-199

ACKNOWLEDGEMENTS

The staff members of Human Factors Research wish to express their sincere appreciation for the continued support given them in their research efforts by the officers and men of the U. S. Fleet Sonar School, San Diego. In particular, we should like to thank the Commanding Officer, Captain C. W. Brigham, and the Executive Officer, Commander J. A. Rinker. We appreciate, too, the assistance given us by Lieutenant Saul Young during the conduct of the experiment reported here.

The study reported here is a part of a larger research project being conducted under contract Nonr 2649(00) with the Office of Naval Research. The research was supported by the Bureau of Ships, the Bureau of Naval Weapons, and the Bureau of Naval Personnel in addition to the Office of Naval Research.

TABLE OF CONTENTS

	Page
CHAPTER I. SUMMARY AND CONCLUSIONS	1
BACKGROUND	1
PURPOSE	2
METHOD	2
CHOICE OF MEASURES	2
SUMMARY OF FINDINGS	4
RECOMMENDATIONS FOR FURTHER RESEARCH	5
CHAPTER II. THE CRITERION MEASURES OF VIGILANCE PERFORMANCE	7
OUTLINE OF THE MAIN STUDY	7
THE CRITERION MEASURES	9
CHAPTER III. THE RELATION OF PSYCHOLOGICAL TEST SCORES TO VIGILANCE PERFORMANCE	13
THE EXPERIMENTAL BATTERY OF TESTS	13
METHOD OF ANALYSIS	19
RESULTS	20
DISCUSSION	20
CHAPTER IV. THE RELATION OF THRESHOLD MEASURES TO VIGILANCE PERFORMANCE	27
METHOD	30
RESULTS	31
DISCUSSION	35
CHAPTER V. RELATION OF SUBJECTIVE REPORTS TO VIGILANCE PERFORMANCE	41
THE DAILY QUESTIONNAIRE	41
PERSONAL INTERVIEWS	48
REFERENCES	53

LIST OF TABLES AND FIGURES

TABLES

		Page
1	Criterion Reliabilities	11
2	Significant Correlations between Tests and Criteria	21
3	Correlations between Percentage of Signals Detected Under Alerted Conditions and Under Watch Conditions	28
4	Relation of Threshold Scores to Vigilance Performance Scores . .	34
5	Reliabilities of Alerted Performance and Threshold Measures and Their Validities for Predicting Performance on Watch	37

FIGURES

1	Group performance on the brightness and loudness discrimination tasks (N = 50)	32
2	Percentage of subjects reporting feelings of tiredness or restlessness prior to going on watch at different hours of the day (each data point N = 288)	44
3	Percentage of subjects reporting feelings of tiredness and restlessness during different weeks of the experiment	45
4	Percentage of signals detected on watch as a function of amount of sleep the night before watch	46
5	Relationship between the response to the question "How do you feel right now?" and detection performance	47
6	Vigilance performance related to expressed attitude toward task .	50
7	Vigilance performance related to task preference	51

Chapter I

SUMMARY AND CONCLUSIONS

BACKGROUND

When human observers are required to maintain attentive watch for prolonged periods of time, they often suffer a loss in detection efficiency as the watch progresses. The magnitude of this loss differs greatly among individual observers. Buckner, Harabedian and McGrath (1960) have demonstrated that individual differences not only accounted for a large proportion of the total variance in signal detection measures, but that these differences were highly reliable both within a single watch period and from one watch period to another. Since individual differences in vigilance performance are large and reliable, they should be predictable. With valid predictors of performance the men best suited for vigilance jobs could be selected for such assignments.

It can readily be demonstrated that selecting expert observers with valid predictor devices would be an effective method of reducing or abolishing the loss of detection efficiency during prolonged watchstanding. It has been repeatedly demonstrated (Holland, 1958; Mackworth, 1950; Solandt and Partridge, 1946) that within any group of observers there is likely to be a subgroup of expert observers who suffer little or no loss of detection efficiency as a function of time on watch. These subgroups have been of considerable size—ranging from 20% to 50% of the total group. Obviously, if such expert observers can be identified beforehand, they can be selected for watchstanding assignments and the less efficient observers can be eliminated. As a result, group performance would be substantially improved.

In spite of the promise of high returns, there has been no systematic attempt to discover or develop instruments for selecting vigilant observers. In view of the increasing requirement for human monitors in military detection systems and in automated industries, the high cost of training operators in such systems, and

particularly the consequences attending the failure to detect certain signals, one can only conclude that such an attempt is long overdue.

PURPOSE

The research described in this report was intended to serve as a starting point in the development of predictors of vigilance performance. The study was exploratory; its purpose was to investigate the relationships between a large number of behavioral measures and criteria of performance on vigilance tasks. The effort was directed toward ascertaining the types of behavioral measures, rather than the specific measurement instruments, that would be promising predictors of vigilance performance. The results of this investigation would then be used to guide further research in the development of specific selection instruments.

METHOD

A study of individual differences in vigilance performance (hereafter referred to as the main study) was conducted concomitantly with the present investigation. The main study has been reported in detail elsewhere (Buckner, et al., 1960) and is briefly outlined in Chapter II of this report. The data collected in that study provided reliable criterion measures of the performances of 54 subjects on an auditory and a visual vigilance task. Many additional behavioral measures were obtained from these subjects prior to, concomitantly with, and following the main study. The relationships between these measures and the criteria of performance were then determined.

CHOICE OF MEASURES

Psychological Tests. In developing a selection program it is usually advisable to explore the possibilities of using psychological tests as selection instruments. Through their use it is possible to obtain objective measures of a wide sampling of human behavior. Tests have the further advantages of being generally economical

and easy to administer. Therefore, most of the measures obtained for this study were taken from "paper-and-pencil" tests. These tests were chosen on the basis of tentative hypotheses about the aptitude, temperament, and motivational variables that seemed to be important in the performance of vigilance tasks. These hypotheses, descriptions of the tests, and the results of the testing program are presented in Chapter III.

Threshold Measures. In the main study it was found that performance under momentary alerted conditions was significantly related to performance under prolonged watchstanding conditions. The variance in the measures of alerted performance appeared to be principally attributable to differences in individual sensory acuities. This led to the hypothesis that measures of thresholds for the required discrimination would be correlated with vigilance performance measures. A detailed discussion of this hypothesis and its subsequent test is contained in Chapter IV.

Subjective Reports. In the main study some of the variance in the vigilance performance measures was designated error variance, because it was associated only with one set of measurements and was not reproduced at other times. Such error variance corresponds to the amount of unreliability in the performance measures. It was anticipated beforehand that a certain amount of such variance in the performance measures could be attributable to temporary characteristics of the subjects (i.e., feelings of tiredness) or to the effects of behavior extraneous to the experiment (i.e., amount of sleep obtained the night before watch). If some of these factors could be identified, some of the error variance could be accounted for, and our general understanding of vigilance phenomena would be advanced.

Data were gathered through subjective reports during the course of the main study and at its conclusion to aid in evaluating the possible sources of unaccountable variance in the measures of vigilance performance. The method of gathering

these data and the results of data analyses are discussed in detail in Chapter V.

SUMMARY OF FINDINGS

1. A number of significant correlations were obtained between psychological test scores and criteria of vigilance performance, but these were generally low in magnitude. None of the psychological tests used in this study was valid enough to be useful by itself in personnel selection. Some had moderate validities, however, and it may be possible to combine certain tests into a battery that will yield a useful multiple correlation with vigilance performance criteria. It is essential first that the present findings be cross-validated on on a new sample of subjects using tasks which more closely approximate actual sonar watchstanding conditions.

2. Tests measuring clerical abilities appeared to be promising predictors of the amount of decrement in detection performance suffered by individuals during watch, but did not appear to predict the overall performance levels.

3. Performance on an auditory vigilance task was more predictable from psychological test scores than performance on a visual vigilance task. Psychological tests may be limited in their usefulness to prediction of performance on auditory vigilance tasks or on tasks that do not demand voluntary attention on the part of the observer.

4. Qualitative differences in vigilance performance (sleeping vs. not sleeping on watch) were more predictable from psychological test scores than quantitative differences in vigilance performance (percentage of signals detected).

5. There was a significant correlation between brightness discrimination threshold and performance on a visual vigilance task. A similar, but nonsignificant, correlation was obtained between loudness discrimination threshold and performance on an auditory vigilance task.

6. There was a steady increase in reports of feelings of tiredness throughout the day. Subjects detected fewer signals when they reported feelings of tiredness

than when they reported feeling rested.

7. Feelings of restlessness increased from week to week during the course of the experiment. These feelings were most pronounced during midmorning and mid-afternoon, but were not related to detection performance.

8. The kind of work (labor, clerical) the subjects performed prior to going on watch had no effect on detection performance, but there was some indication that subjects detected more signals when they came on watch from work that interested them than when they came from boring work.

9. The percentage of signals detected on watch was positively related to the amount of sleep the subject obtained the night before watchstanding. Performance on watch was best when the observer obtained a full night's sleep (8 or 9 hours) and was poorest when the observer had obtained less than two hours sleep. The results suggest that men assigned to vigilance tasks should be allowed sufficient time off watch to obtain adequate sleep.

10. The subjects' general attitudes toward the experiment were not related to their performances on watch. But, there was a significant relationship between task preference (auditory vs. visual) and task performance such that performance was better on the preferred task.

RECOMMENDATIONS FOR FURTHER RESEARCH

Additional research is needed to verify and expand upon the present findings. Fortunately, it should not be necessary to design and conduct experiments exclusively for these purposes. In future investigations of vigilance phenomena in which reliable performance measures are obtained from fairly large groups of subjects, it should be feasible to obtain measurements similar to those used in this study without interfering with the primary purpose of the investigation. That is, it should be possible to incorporate cross-validation and validity generalization studies into research designed primarily for other purposes.

It would be particularly desirable to obtain cross-validation results for four psychological tests that correlated with vigilance performance measures in this study: the MMPI "K" Scale, Attention Test, Visual Speed and Accuracy Test, and O-dotting Test. Tests of clerical skills other than those tested in the present study could also be included.

The research described in this report was conducted using laboratory vigilance tasks whose relationship to actual sonar tasks has not been demonstrated. Validity generalization studies should be conducted in which the watchkeeping task more closely approximates the conditions of sonar watchstanding.

Validity generalization studies would also be important in assessing the importance of discrimination sensitivity in vigilance performance. In this study sensitivity was found to be more important in a visual vigilance task than in an auditory vigilance task. Whether this result was an artifact of some condition in the present study or a general effect could be ascertained by testing the importance of sensitivity in performing other visual and auditory vigilance tasks.

The effect of sleep deprivation on vigilance performance would seem to be a problem for direct experimental investigation. The relationship could be more accurately specified in a controlled experiment. Follow-up studies could then be conducted to investigate the relative merits of different watchstanding schedules for men who perform vigilance tasks.

Chapter II

THE CRITERION MEASURES OF VIGILANCE PERFORMANCE

The data collected in the main study of individual differences in vigilance performance provided the criterion measures used in the present study. These measures will be described in this chapter.

OUTLINE OF THE MAIN STUDY

Subjects. Fifty-four Navy enlisted men assigned to the U. S. Fleet Sonar School, San Diego, participated in the study. At the time the experiment was conducted, these men were waiting for their sonar school classes to convene. They were assigned to participate in the experiment in the same way they were assigned their regular duties. That is, they were told to treat it as an ordinary watch at sea.

All of the subjects were in the upper third of the Navy population in terms of combined Navy General Classification and Arithmetic Reasoning test scores. All had normal near vision and hearing, and all had been in the Navy between six and twelve months at the time of the experiment. The ages of the subjects ranged from 17 to 23 years. All but two had completed high school; all but three were unmarried.

The Watchstanding Task. Two watchstanding tasks were used. In one task, the subjects were required to detect a change in the loudness of a 750 cps tone presented over headphones. The tone was presented intermittently, being on for one second and off for two. In the other task, the subjects were required to detect a change in the brightness of a light appearing in a one-inch square ground-glass covered aperture. The light also was intermittent, on for one second and off for two. Upon detecting a signal (a slight increase in loudness or brightness), the subject pressed a hand-held switch as quickly as possible. This response was recorded automatically by equipment in the experimenter's control room.

The Watch Schedules. Each subject stood a total of 32 one-hour watches, 16 on the visual display and 16 on the auditory. Each subject stood watch for one hour in the morning and one hour in the afternoon, four days a week, for four weeks. The hours of watch were rotated from one week to the next among three randomly selected groups of 18 men each. There was always a five-hour interval between the morning and the afternoon watch. For each subject half of the morning and afternoon watches were on the visual display and half on the auditory display; and half of the watches during each of the two fortnights of the experiment were visual and half were auditory.

Pretests and Posttests. Immediately preceding and following each one-hour watch, the subjects were given a two-minute alerted detection test in which five signals were presented at variable intervals. Auditory tests were given before and after auditory watches, and visual tests were given before and after the visual watches. The signals in the pretests and posttests were of the same intensity (detectable about 90% of the time under alerted conditions) as the signals presented in the hour watches.

Procedure. Each observer stood watch in a separate watchstanding booth. Eighteen subjects were run simultaneously during each watch, nine on the visual and nine on the auditory displays. At the beginning of each watch, a red warning light on a display box in each booth was flashed to indicate the beginning of a warm-up session which lasted until each observer had detected one signal. This session seldom lasted for more than one minute. Then the warning light was flashed to indicate the start of the two-minute pretest. After the pretest, it was flashed again to indicate the start of the one-hour watch. At the end of the watch, the warning light was again flashed to announce the beginning of the two-minute post-test. Following the posttest, the observers returned to their other assigned Navy

duties.

THE CRITERION MEASURES

The response recording apparatus produced a record of the onset and offset of each background stimulus, each signal, and each response made by each subject. From these records, several types of performance measures were obtained for each of the two modes.

Percentage of signals detected. Each observer was presented a total of 384 signals during the 32 one-hour watches, 192 signals on the visual watches and 192 signals on the auditory watches. The percentage of these signals detected by each observer was taken as the major criterion of performance. The reliabilities of these measures, based on the correlation (corrected for double length) between performance scores during the first and second fortnights were .89 for the visual task and .72 for the auditory task.

Latency of response. For each detection, the time interval between the onset of the signal and the onset of the subject's response was measured to the nearest quarter second. The latency of response score indicated the average amount of time the subject took to respond to those signals he detected. High latency scores, since they represented slow reaction times, indicated poor performance. Latencies of false detections (responses indicating the detection of a signal when, in fact, no signal was presented) were not included in the scores, and no time constant was included for missed signals. A signal was considered to be missed if not responded to before the onset of the next background stimulus. This requirement limited the range of latency scores to a maximum of three seconds (the time interval between the onset of a signal and the onset of the next background stimulus). The correlation between the latency of response scores and the percentage of detections was not significantly different from zero for either mode. The correlation between latency scores on the visual watches and latency scores on the auditory watches was .67. The reliability of these scores as estimated by the correlation (corrected for double length) between latency scores during the first and second fortnights was .70 for the visual task and .68 for the auditory task.

Decrement scores. The percentage of detections for the total group declined as a function of time on watch. The amount of decline was different for different subjects. There was an immediate decline in the percentage of signals detected from the pretest to the first part of the watch, and an additional decline during the watch. Since the two decrements may have reflected two different processes, two different decrement scores were derived: pretest-to-watch decrement and within-watch decrement. High decrement scores, like the high latency scores, indicated poor detection proficiency.

The pretest-to-watch decrement score was the difference between the percentage of signals detected under alerted conditions (combined pretest-

posttest scores) and the percentage of signals detected under prolonged watch conditions. The reliability estimates (first vs. second fortnight correlations) for these measures were .26 for the visual task and .77 for the auditory task. Although the reliability of the visual pretest-to-watch decrement score was quite low, it was theoretically possible for other measures to correlate with it up to about .51, and might therefore be useful in an exploratory research program.

The within-watch decrement score was the difference between the percentage of signals detected during the first quarter hour of watch and the percentage of signals detected during the quarter hour in which the maximum loss took place for a particular subject. The reliabilities of the within-watch decrement scores were .53 for the visual task and .52 for the auditory task.

Sleeper vs. non-sleeper. Occasionally, the experimenters discovered subjects sleeping on watch. As an additional criterion of vigilance performance the subjects were divided into two groups: those who had been discovered sleeping on at least one watch and those who had not been discovered sleeping on any watch. It turned out that half of the subjects fell in the sleeper group and half in the non-sleeper group. Of course, the non-sleeper group probably contained some subjects who had slept on watch, but who had not been caught. It was not possible to estimate the reliability of this classification because very few subjects had been discovered sleeping on more than one watch.

False detections. It had been intended that the number of false detections would serve as a criterion measure. After the first day of watchstanding, however, the number of false detections decreased almost to zero. As a result, there was inadequate discrimination between subjects on this measure, and reliable scores could not be assigned. Therefore, the false detection scores were not used as criterion measures.

All of the criterion measures (with the exception of the dichotomous sleeper criterion) yielded normal distributions of scores. The reliabilities of these measures are summarized in Table 1.

Table 1

Criterion Reliabilities

Criterion Measure	Reliability	
	Visual Task	Auditory Task
Percentage of signals detected	.89	.72
Latency of response	.70	.68
Pretest-to-watch decrement	.26	.77
Within-watch decrement	.53	.52
Sleepers vs. non-sleeper	no estimate	

Chapter III

THE RELATION OF PSYCHOLOGICAL TEST SCORES TO VIGILANCE PERFORMANCE

Seventeen different psychological tests yielding 30 separate scores were tried out as possible predictors of vigilance performance. These included tests of intellectual aptitude, temperamental traits, and presumed motivational variables. Some of these had been administered to the subjects as a part of a basic classification battery used by the Navy. The remainder were administered to the subjects ten days after the final watch of the main study, except for one test which was administered once a week during the course of the main study.

THE EXPERIMENTAL BATTERY OF TESTS

The Navy Classification Battery. The scores on the following tests were available from the official service records of the subjects:

1. General Classification Test. (This test may be regarded as a test of general intellectual aptitude.)
2. Arithmetic aptitude.
3. Radio aptitude.
4. Sonar aptitude.
5. Mechanical aptitude.
6. Clerical aptitude.
7. Electronic technician selection test.

Aptitude Tests. To supplement the aptitude tests included in the Navy classification battery, these others were administered ten days after the main study:

8. Visual Speed and Accuracy.

> Performance on vigilance tasks appeared to be similar to performance on simple clerical tasks in that both demanded sustained attention and the detection of small discrepancies,

so the Visual Speed and Accuracy test was included in the battery. This test required a person to compare numbers, symbols, and letters presented in pairs and to indicate whether the items in each pair were identical. The test yielded a speed score (the number of items completed correctly) and an accuracy score (the number of errors made). Hypotheses: (1) The speed score would be positively related to detection proficiency.[1] (2) The error score would be negatively related to detection proficiency.

9. Attention Test.

Wittenborn (1943) reported a factor analytical study of a number of tests designed to measure "attention." The test having the highest factorial validity was included in the present experimental battery. The Attention Test required the subjects to listen to successive series of three digits presented over a loudspeaker. They responded only when the first digit in the series was lowest and the third digit was highest or when the first digit was highest and the second digit was lowest. Two hundred three-digit series were presented in rapid succession after a slower practice session. Approximately one third of these series met the specifications requiring a response. Two scores were derived: (1) A false detection score (the number of incorrect responses) and (2) a total score (the number of correct responses minus one half the number of false detections). Hypotheses: (1) The number of false detections on the Attention Test would be positively related to the number of false detections on the vigilance tasks. Since the false detection score for the vigilance tasks did not prove to be a useful measure, this hypothesis could not be tested. (2) The total score on the Attention Test would be positively related to detection proficiency.

10. Memory Span.

In the vigilance tasks, the subjects were required to compare the intensity of a stimulus with the intensity of stimuli presented earlier. It seemed that immediate memory might play a part in the performance of such tasks. In the Memory Span test, a series of digits was presented over a loudspeaker, and the subjects were required to reproduce the series immediately after the last digit was presented. Forty series, containing from two to twelve digits each, were presented. The score on the test was the mean length of the series that the subject accurately reproduced. Hypothesis: Score on the Memory Span test would be positively related to detection proficiency.

[1] High "detection proficiency" was considered to be reflected by any of the following criteria: high percentage of signals detected, low decrement score, low latency score, or membership in the non-sleeper group.

11. Circle Reasoning.

The items in the Circle Reasoning test consisted of five rows of circles and dashes. One of the circles in each of the first four rows was blackened. The circle was blackened according to some rule or system. The subject's task was to discover the rule and mark the circle that should be blackened in the last row. It was hypothesized that the ability to educe such patterns might be related to the ability to discover the pattern of signal presentation in the vigilance tasks. Since the signal presentation was random, this would lead to the discovery of false patterns and therefore to more false detections. The false detection criterion was not used, however, and the hypothesis could not be tested.

12. Brick Uses.

The Brick Uses test required the subject to list all of the possible uses for a brick that he could think of during a given period of time. This test was developed by Guilford and his associates (1957) to measure flexibility and fluency in thinking. A person listing only construction uses of a brick, for example, would be considered rigid in his thinking as opposed to a person listing the possible uses of a brick as a construction unit, decorative device, weapon, toy, tool, weight, insulator, and so forth. It was felt that a person with more rigid ideation would perform better on vigilance tasks than a person who was more flexible. Hypothesis: Score on the Brick Uses test would be negatively related to detection proficiency.

Temperament Measures. It is reasonable to suppose that temperament variables determine some of the differences among individuals in vigilance performance. Measures of certain temperament traits were included in the experimental battery with the hope that the results would offer some clue to the nature of such determinants.

13. The Guilford-Zimmerman Temperament Survey.

The GZTS is an inventory-type measure of 10 temperament traits that were identified through factor analysis (Guilford and Zimmerman, 1949). Only five of these traits were measured in the present study:

a. General Activity. Baker (1959) and Bowen (1956) found that the more active subjects performed poorer on vigilance tasks than the less active subjects. The General Activity scale might reflect a similar measurement of tendency to

be restless. Hypothesis: Score on the General Activity scale would be negatively related to detection proficiency.

b. Restraint. Persons who score high on the restraint scale are considered to be serious minded, deliberate, and capable of exercising self control. Persons who score low on this scale are considered to be impulsive and carefree. Hypothesis: Score on the Restraint scale would be positively correlated with latency of response on the vigilance task.

c. Ascendance. Persons who score high on this scale are considered to be aggressive and perhaps domineering. Persons who score low are considered to be submissive and in the habit of following rather than leading. It was hypothesized that the more submissive type of person would be more likely to follow the instructions on the rather tedious vigilance tasks, would be less likely to defy the rules by engaging in extraneous activity and, therefore, would perform better than the ascendant type of person. Hypothesis: Score on the Ascendance scale would be negatively related to detection proficiency.

d. Sociability. No hypothesis.

e. Emotional Stability. No hypothesis. Both the Emotional Stability scale and the Sociability scale were included because the scores were available from the answer sheets used for the other three GZTS scales.

14. Manifest Anxiety.

The Manifest Anxiety scale (Taylor, 1953) is derived from certain selected items on the Minnesota Multiphasic Personality Inventory (MMPI). The score obtained from this scale may reflect an individual's generalized drive level. Hypothesis: Score on the Manifest Anxiety scale would be positively related to detection proficiency. The items from two other scales from the MMPI were interspersed among the Manifest Anxiety scale items to act as buffers:

a. "K" scale: a measure of the degree of guardedness or cautiousness with which an individual answers the items on the MMPI.

b. "L" scale: a measure of gross falsification of responses to the items on the MMPI.

15. "Willingness to Guess."

Some individuals may have to be very sure of their detections before responding, while others respond any time they think the stimulus even slightly resembles a signal. This may account for some of the individual differences in vigilance performance. The "Willingness to Guess" test was constructed especially for this study and was intended to measure the degree to which a person is willing to guess when he does not know what the correct response should be. The test was disguised as a test of knowledge of Naval history. Half of the multiple-choice items on the "Naval history" test were legitimate questions of historical facts, most of which were fairly common knowledge. The other half were questions which were entirely fictitious, the correct answers to which could not possibly be known simply because there were no correct answers listed among the alternatives. The subjects were instructed to answer the items that they thought they knew, but to avoid guessing because they would be penalized for incorrect answers. The "Willingness to Guess" score was simply the number of fictitious items that the subject attempted. It was originally hypothesized that this score would be positively correlated with the number of false detections made by the subjects, but the hypothesis could not be tested because the false detection criterion was not used. It also seemed reasonable that this reluctance to guess would be reflected in the latency scores. Hypothesis: Scores on the "Willingness to Guess" test would be negatively correlated with latency of response.

Motivational Variables. It is difficult to measure motivational variables with psychological tests, perhaps because such variables are highly specific to the task to be performed. Nevertheless, an attempt was made to study general motivational factors by using the following two tests:

16. The Behavior Interpretation Inventory (BII).

The Behavior Interpretation Inventory was developed by Applezweig and Moeller (1958) to measure what they considered to be the basic underlying motives of human behavior. The test required the subject to interpret certain behavioral acts, which were verbally described, and to indicate what he thought motivated the act. The test yielded scores on four different scales:

a. Escape. The Escape scale measured the degree to which a person was motivated by a need to escape unpleasant situations. A subject so motivated might be inclined to go to sleep on watch, engage in activities extraneous to the task, or otherwise escape the monotony of

watchkeeping. Hypothesis: Score on the Escape scale would be negatively related to detection proficiency.

b. Avoidance. The Avoidance scale measured the degree to which a person was motivated by a need to avoid future unpleasant situations. Subjects motivated by such a need might be inclined to keep watch more diligently, and thereby avoid the (imagined) consequences of poor performance. Hypothesis: Scores on the Avoidance scale would be positively related to detection proficiency.

c. Social Approval. The Social Approval scale measured general motivation to please others and to be admired by others. No hypothesis.

d. Self Approval. The Self Approval scale measured a person's need to meet self-established standards of achievement. No hypothesis.

17. O-dotting Test.

In the O-dotting test the subject was given a sheet of paper on which were printed many rows of Os. He was instructed to make a dot in the center of as many Os as possible within a given period of time. This task appears to be a finger dexterity task, but there is some evidence (Foy, 1959) that motivation plays an important role in determining performance on it. The O-dotting test may measure the degree to which a person is willing to do what he is told to do, even though the task seems meaningless and is very tiresome and boring. The test was administered on the morning of the second day of each week during the four weeks of the main study. The subjects were told nothing about the purpose, meaning, or use of the test. They dotted Os continuously for four minutes, marking their places at the end of each minute. After a 40-second rest, they dotted Os for another minute. Three different scores were obtained from the test results:

a. Total score: The total score was the mean number of Os dotted in the five-minute periods. Hypothesis: Total score on the O-dotting test would be positively related to detection proficiency.

b. Decrement score: The decrement score was the mean difference between the number of Os dotted in the first two minutes and the number of Os dotted in the third and fourth minutes. Hypothesis: The decrement score on the O-dotting test would be negatively related to detection proficiency.

c. Recovery score: The recovery score was the difference between the mean number of Os dotted in the first four minutes and the number dotted

in the fifth minute (after the rest period). The recovery score could be considered to represent the discrepancy between the subject's performance during continuous work and performance during a brief "end spurt." The smaller the discrepancy, the closer to capacity will be the subject's work rate. Hypothesis: The recovery score would be negatively related to detection proficiency.

METHOD OF ANALYSIS

Complete test results could not be obtained on five subjects, so these subjects were eliminated from the first phase of the analysis. A frequency distribution of the scores of each of the 49 remaining subjects on each test was tabulated. The scores were then transformed to a 7-point normalized scale so that the frequency distributions approximated the following:

Coded Score:	1	2	3	4	5	6	7
Frequency:	2	5	10	15	10	5	2

The criterion scores described in Chapter III were similarly transformed to the 7-point scale. To recapitulate, these criterion scores were:

1. Percentage of signals detected on all visual tasks.
2. Percentage of signals detected on all auditory tasks.
3. Pretest-to-watch decrement on all visual tasks.
4. Pretest-to-watch decrement on all auditory tasks.
5. Within-watch decrement on all visual tasks.
6. Within-watch decrement on all auditory tasks.
7. Mean latency of response on all visual tasks.
8. Mean latency of response on all auditory tasks.
9. Sleeper vs. non-sleeper.

The test and criterion scores in coded form were punched on IBM cards and through a process of sorting and counting on the IBM 101 statistical machine, a scatter diagram was obtained representing the relationship of each test score to each criterion. The scatter diagrams were inspected and those that indicated a zero or near zero relationship were eliminated. When a possibly significant relationship was indicated, the scatter diagram was inspected for linearity and

homoscedasticity. The appropriate coefficient of correlation (biserial r's for the sleeper criterion, product moment r's for all others since the assumptions for these coefficients were met in all cases) between the two variables was computed from the original uncoded scores using all available subjects.

RESULTS

The significant correlations found between tests and criteria are reported in Table 2. Only those coefficients of correlation that were different from zero at the .05 level of significance are reported. (It should be recalled that higher decrement scores indicate lower proficiency, and therefore, the correlations between tests and decrement scores must be reversed in sign to indicate the relationship between the test score and detection proficiency.)

There were no significant correlations with the latency of response scores. Out of 90 correlations with the visual detection criteria (percentage detection and decrement scores), 5 were significant at the .05 level and none at the .01 level. Out of 90 correlations with the auditory detection criteria, 14 were significant at the .05 level and 4 of these were significant at the .01 level. Out of 30 correlations with the sleeper criterion, 6 were significant at the .05 level and 3 of these were significant at the .01 level. All together, 17 measures correlated significantly with one or more criteria and 13 measures had no significant correlations with any criterion. The highest correlation obtained (-.49) was between the "K" scale scores and percent auditory detections.

DISCUSSION

The possible interpretations of the results presented in Table 2 are limited. Because of the large number of correlations computed, it would be expected that many of them would by chance alone exceed the criterion adopted for statistical significance. It is possible, for example, that none of the "significant" correlations

Table 2

Significant Correlations between Tests and Criteria

Task	Criterion	Test	Correlation between Test and Criterion
VISUAL	Percent detections	Electronic Technician Selection Test	.32
		O-dotting (recovery)	-.29
		Mechanical Aptitude	.29
	Pretest-to-watch decrement	Mechanical Aptitude	-.31
	Within-watch decrement	Visual Speed & Accuracy (speed)	-.31
AUDITORY	Percent detections	MMPI "K" Scale	-.49*
		Sonar Aptitude	.34
		GZTS Ascendance	-.32
		BII Avoidance	.29
		O-dotting (recovery)	-.29
	Pretest-to-watch decrement	GZTS General Activity	.41*
		O-dotting (recovery)	.34
		GZTS Ascendance	.32
		BII Avoidance	-.30
	Within-watch decrement	Attention Test (total score)	-.47*
		Visual Speed & Accuracy (error score)	.39*
		Clerical Aptitude	-.33
		Attention Test (false signals)	.31
		Brick Uses	.29
VISUAL AND AUDITORY	Sleeping on at least one watch vs. not sleeping on any watch	Visual Speed & Accuracy (error score)	-.44*
		Mechanical Aptitude	-.44*
		O-dotting (total score)	.39*
		O-dotting (decrement score)	-.38
		Brick Uses	.38
		BII Escape Scale	-.31

* Different from zero at the .01 level of significance. All other coefficients are significant at the .05 level.

with the visual detection criteria (5 out of 90) represents a real relationship outside of the particular sample used in this study. To estimate the actual number of significant correlations to be expected by chance, the intercorrelations of all the tests included in the battery must be known. This information was not obtained in the present study. If the tests were all independent, 5 out of 100 correlations

would be expected by chance to be significant at the .05 level. Since the scores on the tests were probably intercorrelated to some extent, it would be expected that less than 5 would be significant by chance.

It would seem reasonable to conclude that many of the correlations with the auditory detection criteria indicate true relationships outside of the present sample, because the number of significant correlations (14 out of 90) most certainly exceeds what would be expected by chance. Nevertheless, it must be recognized that some of the correlations reported in Table 2 undoubtedly represent mere correlated errors of measurement or chance occurrences, and these results must be interpreted with caution. It would be unwise to conclude that any of the reported correlations indicates validity for selecting vigilant performers without first cross-validating the results on a new sample of subjects.

A further restriction in the obtained correlations was caused by preselection of the subjects on certain tests of the Navy battery. The subjects were an exceedingly homogeneous group. They had been selected for Sonar School on the basis of some of the tests of the Navy battery which had been found to be correlated with success in that school. This preselection restricted the range of scores on the tests of the Navy battery and other tests with which they are correlated. This restriction in range would reduce, perhaps considerably, the possible validity of those scores for predicting performance on the vigilance tasks. Therefore, the fact that few of the tests of the Navy battery produced significant correlations with the criteria of vigilance performance should not be taken as clear evidence of the lack of efficiency of these tests in a selection program for watchstanders.

Although none of the obtained correlations appeared to be large enough for any single test to be useful in personnel selection, many correlations were of moderate size and certain tests could possibly be combined into a battery to yield a useful multiple regression equation for predicting vigilance performance. It would not be profitable to develop such a multiple regression equation unless cross-validiation results have been obtained.

The results supported the following hypotheses:

1. The speed score on the Visual Speed and Accuracy test would be positively related to detection proficiency (-.31 with visual within-watch decrement).[2]

2. The error score on the Visual Speed and Accuracy test would be negatively related to detection proficiency (.39 with auditory within-watch decrement).

3. The total score on the Attention Test would be positively related to detection proficiency (-.47 with auditory within-watch decrement).

4. Score on the Brick Uses test would be negatively related to detection proficiency (.29 with auditory within-watch decrement).

5. Score on the GZTS General Activity scale would be negatively related to detection proficiency (.41 with auditory pretest-to-watch decrement).

6. Score on the GZTS Ascendance scale would be negatively related to detection proficiency (.32 with auditory pretest-to-watch decrement; -.32 with percentage auditory detections).

7. Score on the BII Escape scale would be negatively related to detection proficiency (-.31 with the sleeper criterion).

8. Score on the BII Avoidance scale would be positively related to detection proficiency (.29 with percent auditory detections; -.30 with auditory pretest-to-watch decrement).

9. Total score on the O-dotting test would be positively related to detection proficiency (.39 with the sleeper criterion).

10. The decrement score on the O-dotting test would be negatively related to detection proficiency (-.38 with the sleeper criterion).

11. The recovery score on the O-dotting test would be negatively related to detection proficiency (-.29 with percentage visual detections; -.29 with percentage auditory detections; .34 with auditory pretest-to-watch decrement).

The results did not support the hypotheses concerning the following tests: Memory Span, GZTS Restraint, Manifest Anxiety and "Willingness to Guess."

The results presented in Table 2 indicate that auditory vigilance performance was more predictable from psychological test scores than visual vigilance performance. It will also be noted that, with the single exception of the O-dotting recovery score, no test variable predicted performance on both the visual and auditory tasks. In the main study it was found that there was a low correlation

[2] A positive relationship to detection proficiency would be indicated by a negative correlation with decrement scores.

(r = .30) between performance on the auditory task and performance on the visual task. On the basis of this low correlation and the differences in predictability of performance under the two different modes, it can be concluded that, in spite of the many apparent similarities, the tasks were quite different in the demands they placed upon the subjects. It might be profitable, therefore, to examine closely the differences in the demands of auditory and visual vigilance tasks. If these were more clearly understood, the possibilities of developing predictor devices might be improved.

It is possible that one of the important ways in which the two tasks differed was in the availability of stimuli to the subject. On the auditory task, the stimuli were presented directly to the sense organ through the headphones, but on the visual task, the subject had to attend voluntarily to the display to receive the stimuli emitted by the display. When the subject was monitoring the visual display, the amount of extraneous, non-task-oriented activity in which he could engage and still perform his watchstanding duties was limited by the need to continuously attend to the display. That is, he could not remove his gaze from the display box and still receive the signals. On the other hand, when the subject monitored the auditory display he was much less restricted. Since he would continue to receive the auditory stimuli regardless of what he happened to be doing or looking at, he was free to engage in activities extraneous to the vigilance task (i.e., reading, writing). In interviews conducted after the main experiment many subjects admitted that they occasionally read books on watch and sometimes wrote letters or doodled on scratch pads, even though these activities had been prohibited. They reported that these things were done mainly on the auditory watches and seldom on the visual watches. The implication is that the factors that determine or influence the degree to which subjects differ in amount of extraneous activity may have been measured by some of the psychological tests in the experimental battery. If this were so, then these tests would be effective in predicting

performance tasks in which there are great differences among the subjects in amount of extraneous activity, but would not predict performance on the tasks in which there is little or no difference among subjects in the amount of such activity.

The relationship between within-watch decrement scores and scores on clerical-type tests was fairly consistent. Practically all of the tests correlating significantly with the within-watch decrement criterion required the subject to perform some very routine act repeatedly and rapidly. This uniformity of relationships was found only for the within-watch decrement and not for the other detection criteria. It is possible, in other words, that clerical aptitude is correlated with the degree to which an observer can sustain his performance level throughout a prolonged watch, but is unrelated to the absolute level of that performance. The absolute level of performance evidently would be determined by factors other than clerical aptitude, such as sensory acuity and task difficulty. If predictors of overall level of performance can be developed, clerical aptitude tests may be useful in predicting the degree to which those performance levels will be maintained throughout a prolonged watch.

The sleeper vs. non-sleeper criterion appeared to be the most predictable one. This suggests that differences in vigilance performance that reflect differences in <u>kinds</u> of behavior may be more predictable from psychological test scores than differences in <u>amount</u> of proficiency. That is, the sleeper criterion reflected qualitative differences in performance, while the other criteria reflected quantitative differences. Psychological tests of the type used in this study may be better predictors of qualitative than quantitative differences in vigilance performance.

Chapter IV

THE RELATION OF THRESHOLD MEASURES TO VIGILANCE PERFORMANCE

In the main study it was found that performance under alerted conditions (percentage detections on all pretests and posttests) was correlated with performance under prolonged watchstanding conditions (percentage detections on all one-hour watches). The correlation was .74 for the visual watches and .67 for the auditory watches. These findings were consistent with the results of other studies (Bakan, 1955; Jenkins, 1958; Solandt and Partridge, 1946) in which initial performance levels were correlated with the amount of decline on watch.

The variance in the measures of alerted performance appeared to be largely attributable to individual differences in brightness or loudness difference thresholds. This led to the hypothesis that performance on a vigilance task would be a function of the individual's sensitivity for the discrimination required for detecting a signal on that task.

If the alerted performance measures could be interpreted as measures of threshold sensitivity, then the hypothesis would need little further verification. The correlations of .74 and .67 could then be regarded as validity coefficients for predicting detection performance on the visual and auditory tasks from threshold measures. Such an interpretation could not be made immediately because of two possible sources of confounding. The correlations between alerted and watch performance may have been spuriously high because of the common influence on these measures of transient changes in the subjects' behavior or moods. A second problem was that the alerted performance measures may not have been a clear measure of discrimination threshold because of the searching component in performance on the pretests and posttests.

The measures of alerted performance and watch performance were taken in close temporal order. If there were factors which varied from day to day (i.e., amount

of sleep the night before, feelings of tiredness, etc.) and which influenced performance under both conditions and if there were reliable individual differences in these factors, then the alerted and watch performance measures would tend to rise or fall together--they would be correlated. This would result in the validity coefficients being spuriously inflated. That is, the variance in the measures attributable to these extraneous factors would enter the validity coefficients as systematic variance, when in fact it should be regarded as error variance.

Fortunately, it was possible to ascertain whether or not alerted performance was correlated with watch performance as a result of the influence of such transient factors. This was done by correlating alerted performance during the first fortnight of the experiment with watch performance during the second fortnight and then correcting the resulting validity coefficient for increased (doubled) length in both measures.[1] The results were then cross-validated by correlating alerted performance during the second fortnight with watch performance during the first fortnight. It can be seen from the results reported in Table 3 that when the

Table 3

Correlations between Percentage of Signals Detected Under Alerted Conditions and Under Watch Conditions

	Uncorrected		Corrected for Doubled Length	
	Visual	Auditory	Visual	Auditory
All pre-posttests vs. all watches	.74	.67	--	--
First fortnight pre-posttests vs. second fortnight watches	.59	.49	.76	.60
Second fortnight pre-posttests vs. first fortnight watches	.41	.48	.53	.60

[1] The formula for this correction is given by Guilford (1954, p. 407).

effects of extraneous variables are controlled by correlating alerted performance during one fortnight with watch performance during another fortnight, the resulting correlation coefficients, when corrected for double length, are comparable to the original validity coefficients obtained from all pretests and posttests and all watches. It is possible that correlated errors of measurement increased the original coefficients slightly, but it is clear that their influence does not account for the high correlation between alerted and watch performance.

The second problem in interpreting the correlation between alerted and watch performance as a validity index of the prediction of vigilance performance from threshold measures concerned the influence of the searching component in the pretests and posttests. Even though the pretests and posttests were only two minutes in duration, the observer was required to select from many stimuli those few stimuli which were signals. Therefore, the pretests and posttests may be regarded as brief watchstanding tasks with very high signal rates. The correlation between alerted and watch performance would then more appropriately be considered a reliability coefficient, rather than a validity index.

To test the hypothesis that discrimination threshold is related to vigilance performance, it would be necessary to obtain threshold measurements in a way that minimizes the searching or watchkeeping aspect of the measurement procedure. The measurement procedure should require immediate judgments of differences between pairs of stimuli, rather than require that specified stimuli be selected out of a background of neutral stimuli. Further, the subjects should be allowed frequent rest periods and should not be required to work continuously for more than a few minutes at a time. Therefore, a separate study was conducted in which threshold measurements were obtained under such conditions. Two separate, parallel studies were concurrently conducted: brightness discrimination threshold and performance on the visual vigilance task; loudness discrimination threshold and performance on the auditory vigilance task.

METHOD

Display. The displays used in the threshold measurement study were the same as those used in the main experiment. For the brightness discrimination measures, the intermittent light was used; for the loudness discrimination measures, the intermittent tone was used.

Stimuli. For each mode, standard and comparison stimuli were used. The standard stimulus for the brightness discrimination task was a light of the same brightness as the background stimuli in the visual vigilance task. The standard stimulus for the loudness discrimination task was a tone of the same loudness as the background stimuli in the auditory vigilance task. For both the brightness and loudness studies, ten different comparison stimuli were used. Each comparison stimulus was progressively more intense (brighter or louder) than the corresponding standard stimulus. The intensity levels were set in a pilot study, using HFR office and staff personnel, so that comparison stimulus levels would bracket a range from 100% correct discriminations to 50% correct (chance) discriminations. Both the standard and the comparison stimuli were one second in duration.

Stimulus Presentation. The standard stimulus was paired 20 times with each of the 10 comparison stimuli, making a total of 200 required discriminations on each mode. Stimuli were presented in pairs with a 2-second interval between the stimuli in each pair and a 5-second interval between separate pairs. The standard stimulus appeared in each pair. It was first 50% of the time, and second 50% of the time.

The pairs were randomly presented in groups of 20 pairs. Within each group, every comparison stimulus appeared twice, and was the first member of the pair one time, and the second member the other time. The presentation of 20 pairs took about three minutes.

On the first day, five groups of 20 pairs were presented on one mode. The subjects were allowed a two-minute rest period between each group. Then after a half-hour rest period, five groups of 20 pairs were presented on the other mode. On the second day, the same procedure was used, except the order of presentation on the two different modes was reversed for each subject.

The Subjects' Task. The subjects were required to judge the brightness or loudness differences between each pair of stimuli. They reported each judgment by marking on an answer sheet whether the second stimulus was more intense or less intense (brighter vs. dimmer or louder vs. softer) than the first stimulus in each pair.

Testing Conditions. The brightness discrimination tests were conducted in the watchstanding booths used in the main experiment. The loudness discrimination tests were conducted in a large, sound-controlled room in which the subjects sat apart from each other and with their backs to each other. In both the brightness and the loudness discrimination tests, an experimenter was present in the room and called off the item numbers to the subjects.

RESULTS

Group Performance. The percentage of correct discriminations between the standard stimulus and each comparison stimulus for all subjects is shown in Figure 1. The smooth curves were fitted by inspection. (In the case of the loudness discrimination curve, the data point for the comparison stimulus value of 2 was ignored. It was found that a defect in the recording equipment had produced a louder tone than had been intended for this value.) The two curves indicated that the subjects, as a group, discriminated between the standard and comparison stimuli with increasing accuracy as the intensity of the comparison stimulus—and thus the difference between the standard and comparison stimuli—was increased.

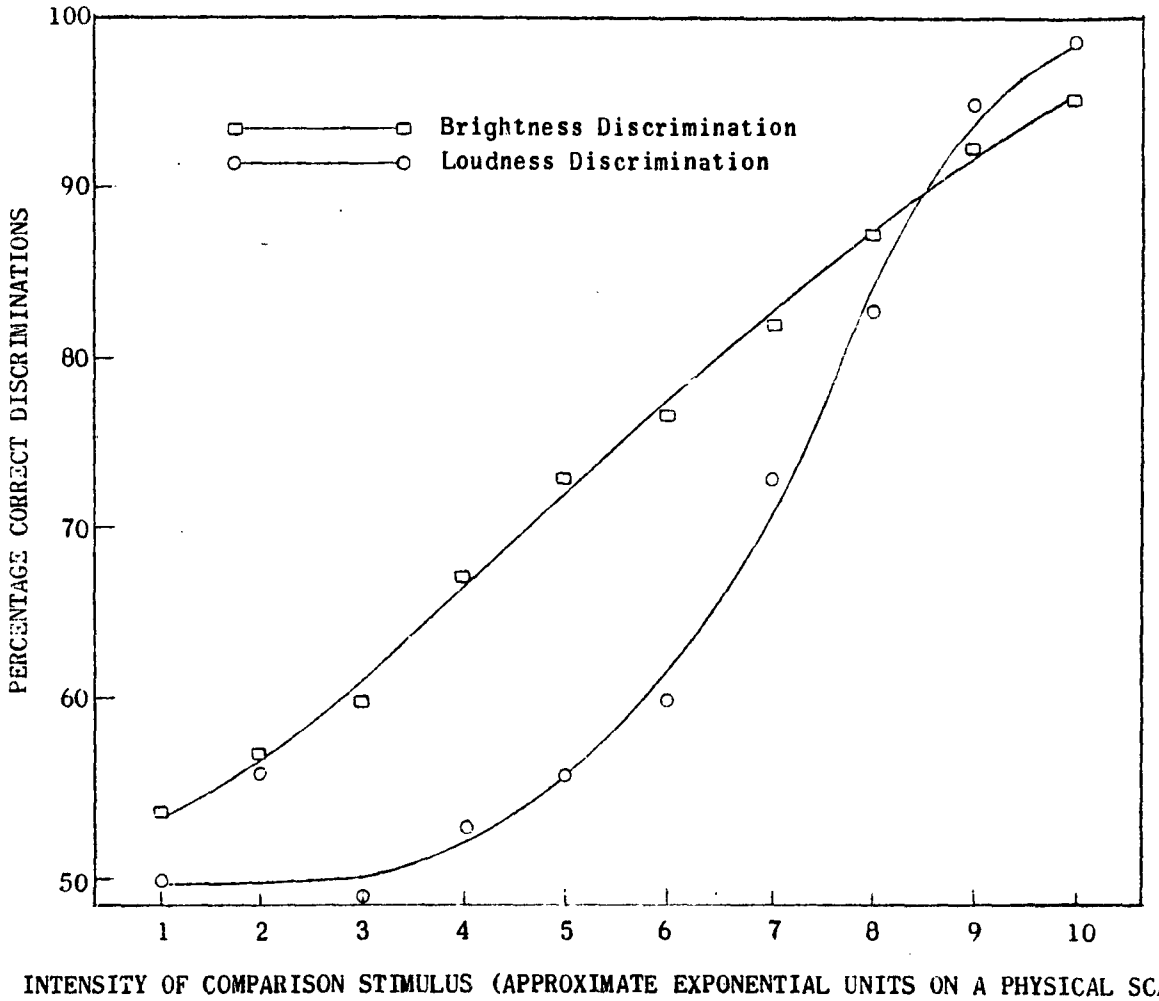

Figure 1. Group performance on the brightness and loudness discrimination tasks (N = 50).

The Threshold Measures. Difference limens (DLs) for loudness and brightness were obtained by plotting each subject's performance in the manner illustrated in Figure 1. The data points for the comparison stimulus values of 1, 2, and 3 were not used in ascertaining individual DLs, because discriminations at those levels involved a good deal of guessing for most subjects, and chance successes or failures would be more likely to influence the location of these data points. A smooth curve was fitted by inspection to the remaining data points. The point on the stimulus dimension which intersected the 75% correct point on the response dimension was taken as the measure of discrimination threshold (the DL).

For the group brightness discrimination curve shown in Figure 1, the threshold score was 5.6; the mean threshold score of the individual brightness discrimination curves was 5.7. For the group loudness discrimination curve, the threshold score was 7.3; the mean threshold score of the individual loudness discrimination curves was 7.2. The close agreement of these results indicated that no systematic bias was introduced by eliminating the data points for the comparison stimulus values of 1, 2, and 3.

The threshold scores were normally distributed for both modes, but there was considerably less variability in the loudness DLs (σ = .45) than in the brightness DLs (σ = 1.15).

Reliability. To obtain an estimate of the reliability of the individual DLs, the data obtained on the first and second days were plotted separately. With the partial data, however, one wrong discrimination meant a shift of 10% on the response dimension; therefore, the data points were too unstable to permit fitting curves by inspection. A different type of score was used to obtain an estimate of the reliability of individual differences in performance on the discrimination tasks. This new measure was simply the total percentages of correct discriminations (eliminating comparison stimulus values 1, 2, and 3) made on the two different days.

The distributions of scores for the two different types of scores (DLs and percentage correct) were practically identical. The percent correct scores correlated -.81 with DLs for both the loudness and the brightness discrimination tasks. The two measures, therefore, were reasonably equivalent.

The test-retest reliabilities of the percentage correct scores were .42 for loudness discrimination and .30 for brightness discrimination. When corrected for double length, these reliability coefficients were .59 and .42. It is probable that the DLs had similar reliabilities.

Validity. The hypothesis that measures of discrimination threshold would be correlated with vigilance performance was tested by correlating DLs with the percentages of signals detected under alerted and watchstanding conditions. The results are presented in Table 4.

Table 4

Relation of Threshold Scores to Vigilance Performance Scores

	Vigilance Performance Score (Percentage Detections)			
	Alerted Conditions		Watch Conditions	
Threshold Score	Visual	Auditory	Visual	Auditory
Brightness DL	-.36*	---	-.41*	---
Loudness DL	---	-.23	---	-.19

* Significant at .01 level.

Brightness DL was significantly correlated with the percentage of signals detected on the visual vigilance task under either alerted or watch conditions. But loudness DL was not significantly correlated with auditory performance under either condition, although the relationship was in the expected direction.

DISCUSSION

The original hypothesis was that performance on a vigilance task would be a function of the individual's sensitivity for the discrimination required for detecting a signal on that task. The visual vigilance task required brightness discriminations and the auditory vigilance task required loudness discriminations. Therefore, two specific hypotheses were made: (1) Brightness DL would be significantly correlated with detection performance on the visual vigilance task. (2) Loudness DL would be significantly correlated with detection performance on the auditory vigilance task. The first hypothesis was supported by the results; the second was not.

Validity of Brightness DLs vs. Validity of Loudness DLs. There are two possible ways of accounting for the greater validity of the brightness DLs for predicting visual vigilance performance compared with the validity of the loudness DLs for predicting auditory vigilance performance. These alternatives concern (1) the variability of the DLs, and (2) the characteristics of the watchstanding tasks.

The loudness DLs were less variable (σ = .45) than the brightness DLs (σ = 1.15). The loudness DLs were thus more restricted in range and less discriminating. This restriction would serve to reduce the possible predictive value of the loudness DLs.

The characteristics of the auditory and visual watchkeeping tasks were similar in most ways, but there was at least one important difference. In the auditory task, the stimuli were presented directly to the subject through headphones, regardless of whether he was voluntarily attending to the stimuli. In the visual task, it was necessary that the subject look at the display to receive the stimuli, and thus voluntary attentiveness was required. In the previous chapter, this difference was discussed in relation to the degree to which performance on the two different tasks could be predicted from psychological test scores. It was suggested

that differences in performance on the auditory task may have been determined to some extent by differences in the amount of extraneous activity engaged in by the subjects while on watch. The incidence of such activity was more common on the auditory watches than on the visual watches, because the auditory task demanded less voluntary attention. If we assume that sensory thresholds are unrelated to amount of extraneous activity and that some of the variance in the detection performance measures was determined by differences in amount of extraneous activity, then sensory thresholds would probably be less related to detection performance on the auditory task than to detection performance on the visual task.

Validity of DLs vs. Validity of Alerted Performance Scores. The threshold measures were not as highly correlated with performance on watch as were the measures of performance on the pretests and posttests. There are several possible reasons for the low validity of the DLs compared with the validity of the alerted performance scores. These reasons involve the possible effects of (1) reliability, (2) training, (3) attitudes, and (4) the characteristics of the measurement procedures.

When comparing the validities of different measures for predicting a criterion, one must consider the reliabilities of those measures. In Table 5 a comparison is shown between the reliabilities of the measures and their validities for predicting performance on watch. It can be seen that within each mode higher validity goes with higher reliability. But, when reliabilities and validities are compared across modes, the relationship does not hold consistently. It must be recognized that the estimate of reliability for the threshold measures was made from a different, although nearly equivalent, score than the one used to estimate the validity of the threshold measures. It may have been that the DLs were considerably less reliable than the percentage correct scores. Thus, it is difficult to evaluate the explanation of the differences in validities of the alerted performance

Table 5

Reliabilities of Alerted Performance and Threshold Measures and Their Validities for Predicting Performance on Watch

	Alerted Performance Scores		Threshold Measures	
	Visual	Auditory	Visual	Auditory
Reliability	.50	.80	.42	.59
Validity	.74	.67	.41	.19

scores and the threshold scores in terms of differences in reliabilities. It remains a possible, but not an unequivocal, explanation.

The threshold measures were taken about a week after the conclusion of the main experiment. The alerted performance measures were taken, of course, during the course of the experiment. The differences in time of measurement could have allowed two other factors to influence the differences in validities: training effects and attitude effects.

It was possible that after four weeks of watchstanding the subjects became well-trained in making the required brightness and loudness discriminations. The subjects may have been a more homogeneous group with respect to brightness and loudness discrimination abilities at the time the threshold measures were obtained than they were at the time the alerted performance measures were obtained. As a group becomes more homogeneous with respect to a particular behavior, measures of that behavior become less valid as predictors. Thus the threshold measures would have been less valid than the alerted performance measures.

It was also possible that the subjects had a different attitude toward the threshold measurement task than they had toward the watchstanding tasks. During the course of the main experiment the subjects were waiting for their Sonar School classes to convene, and when they were not participating in the experiment

they were assigned to odd jobs around the Sonar School. When the threshold measures were taken, however, many of the subjects had begun their training, and were temporarily withdrawn from certain classes to participate in the threshold measurement experiment. It is possible that the subjects did not mind participating in the main experiment, but some may have resented the requirement that they participate in the threshold measurement experiment, because it interrupted their training. In that case, differences in performance on the threshold measurement task may to some extent reflect differences in attitudes, and thereby introduce a source of variance that was irrelevant to their previous watchstanding performance.

A fourth possible source of the different validities of the DLs and the alerted performance measures concerns the characteristics of the measurement procedure. It was mentioned earlier in this chapter that the alerted conditions shared with the prolonged watch conditions a common requirement that the subject select out of several sequentially presented stimuli those which were signals. The common searching element may have been a source of correlation between the alerted and prolonged watch performances. This common element was not shared by the threshold measurement procedure, since the procedure called for momentary comparisons of pairs of stimuli. Therefore, the alerted performance measures would correlate more highly with watch performance than would the threshold measures.

The correlation between alerted and watch performance could not be interpreted as an index of the relationship between discrimination threshold and vigilance performance because this correlation could have been inflated by common task elements. It was the object of the threshold measurement study to assess the amount of this inflation. The results indicated that the relationship between discrimination threshold and detection performance is probably not as great as that indicated by the correlation between alerted and watch performance. It is more likely that the relationship is real, but of a moderately low order. It is also

possible that the importance of discrimination sensitivity is greater on a visual watchstanding task than on an auditory watchstanding task.

Chapter V

RELATION OF SUBJECTIVE REPORTS TO VIGILANCE-PERFORMANCE

During and following the main experiment subjective reports were obtained through questionnaires and interviews. It was anticipated that these data would provide a means of evaluating (1) the influence of the subjects' attitudes on their vigilance performance, (2) any change or shift in attitudes during the course of the experiment, and (3) the effect of various extra-experimental activities on vigilance performance. This information would be useful in ascertaining some of the sources of unaccountable variance in the vigilance performance measures.

THE DAILY QUESTIONNAIRE

Before every watch each subject filled out a brief questionnaire of the check-list type. Each subject completed the questionnaire a total of 32 times during the course of the main experiment (twice a day for 16 days). The questions were designed primarily to obtain information on the general activation level of the subjects at the time they went on watch. It was felt that general activation level might be reflected or influenced by previous activities or by immediate mood. Answers to the following questions were obtained:

1. How long has it been since your last meal? _____ hours.
2. How long has it been since your last liberty? _____ days.
3. How many hours did you sleep last night? _____ hours.
4. Have you been working today? _____ yes. _____ no.
5. (if "yes" to 4.) What kind of work?
   _____ clerical or study.
   _____ light labor.
   _____ heavy labor.
6. (if "yes" to 4.) How did the work interest you?
   _____ very interesting.
   _____ somewhat interesting.
   _____ boring, tiresome.

7. How do you feel right now?
   _____ rested and refreshed.
   _____ fairly rested.
   _____ somewhat tired.
   _____ very tired.

8. What kind of a mood are you in?
   _____ restless and nervous.
   _____ somewhat restless.
   _____ fairly relaxed.
   _____ very relaxed.

Method of Analysis. The responses to the questionnaire items and the percentage of signals detected on the corresponding watch were coded and punched on IBM cards. Three analyses were made:

1. Response trends by time of day. The IBM cards were sorted according to the hour of the day during which the watches were stood. The percentage of subjects responding to each answer category at the different hours was then tabulated.

2. Response trends from week to week. The cards were sorted according to the week during which the watches were stood. Again the percentage of subjects responding to each category was tabulated.

3. Responses and watchstanding performance. To relate the subjects' responses to their subsequent watchstanding performance, the cards were sorted on each question according to response categories. Mean watch performance scores were then obtained for each response category.

In the third analysis all subjects were not represented in all categories an equal number of times, because some subjects responded more often in a particular category than did others. The assumptions underlying the usual tests of statistical significance could not be met. To conduct the tests of significance the cards were sorted for each question in the following way: for each subject two cards were randomly selected for each response category. (Sometimes the categories had to be combined to make this possible.) With question 5, for example, for each subject two cards were randomly selected from among those on which the subject responded "clerical or study," two more were selected from those on which he responded "light labor" and two more for "heavy labor." These selected cards produced matched groups

for each response category—each subject was represented twice in each category. The mean watch scores were then determined for each category and analyses of variance were conducted to test the significance of the differences between means.

Questions 1, 2, and 4 were eliminated from the analysis because the subjects did not respond differentially to these questions. The subjects led a fairly regimented life at the time of the experiment and their mealtimes, work assignments, and days of liberty were very much the same for all subjects. For example, since all subjects took their meals at about the same time each day, the "time since last meal" data merely reflected the time of day.

Results. The response trends during the different hours of the day for questions 7 and 8 are shown in Figure 2. The response categories of question 7 were dichotomized into rested vs. tired, and the categories of question 8 were dichotomized into restless vs. relaxed. The response trends shown in Figure 2 indicate that the percentage of subjects reporting feelings of tiredness increased during the day. Reports of restlessness were most pronounced during midmorning and midafternoon.

There was a slight increase in reports of "restless" as the experiment progressed from week to week (Figure 3). The incidence of reports of "tired" was irregular from week to week but was generally lower for the second fortnight. Except for the first week, there appeared to be a steady increase in feelings of tiredness from day to day within each week.

There was a marked relationship between the amount of sleep reported (question 3) and percentage of signals detected on watch. Subjects performed better on watch when they had several hours sleep the night before than when they had only a few hours sleep. This relationship is shown in Figure 4, and was found to be statistically reliable ($p < .05$).

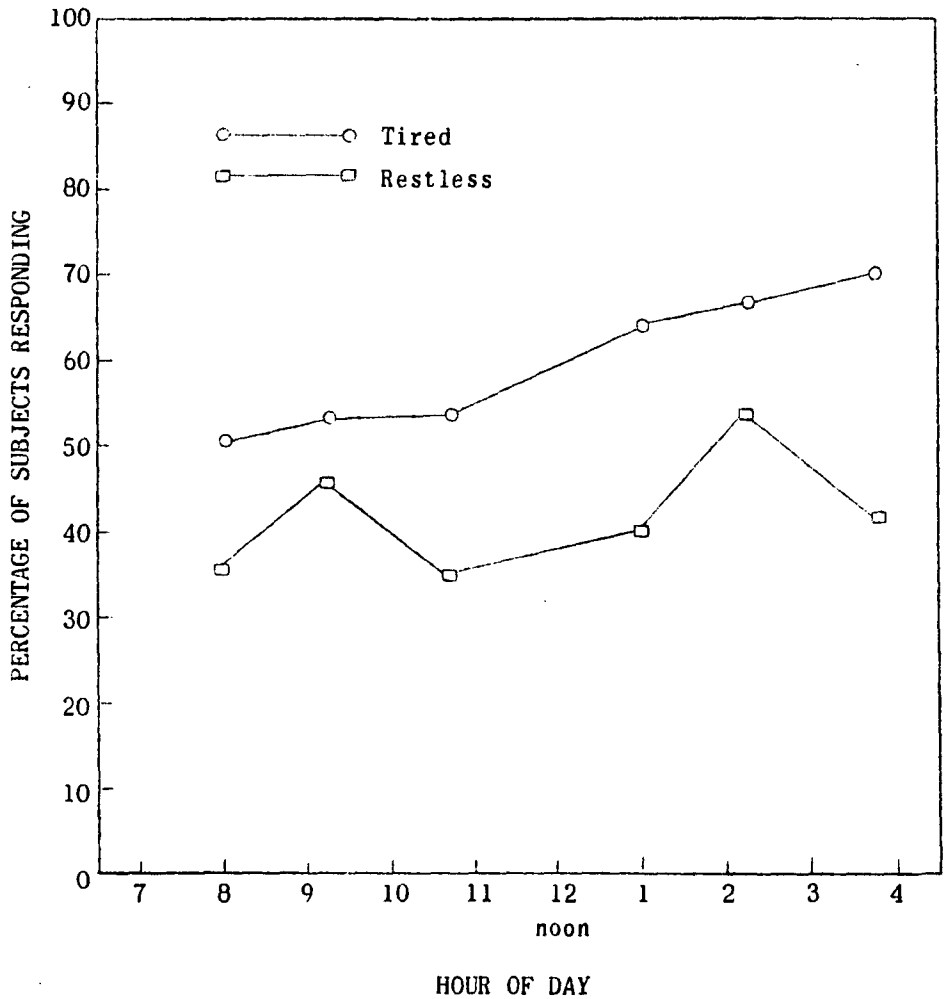

Figure 2. Percentage of subjects reporting feelings of tiredness or restlessness prior to going on watch at different hours of the day (each data point N = 288).

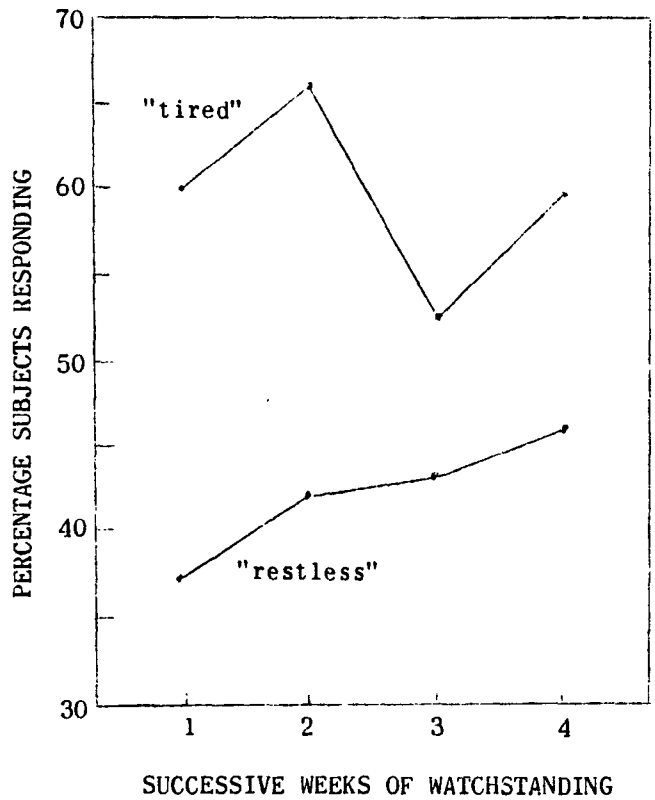

Figure 3. Percentage of subjects reporting feelings of tiredness and restlessness during different weeks of the experiment.

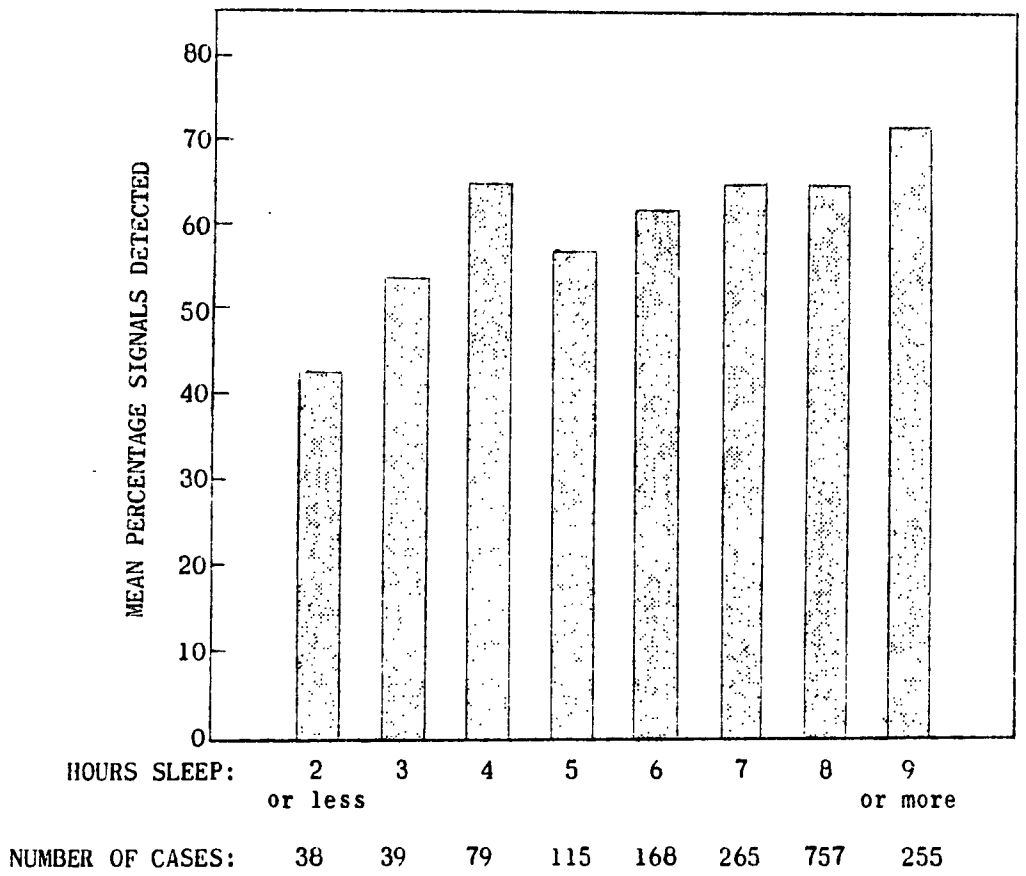

Figure 4. Percentage of signals detected on watch as a function of amount of sleep the night before watch.

There was no relationship between the kind of work (question 5) performed prior to coming on watch and detection performance. There was some indication from the responses to question 6, though, that subjects performed slightly better when they came from work that interested them (67.1% detections) than when they came from boring work (62.7% detections).

Subjects performed poorer on watch when they felt tired than when they felt rested (Figure 5). The data from the response categories "fairly rested" and "rested and refreshed" were combined in Figure 5 because there were very few cases in the latter category.

There appeared to be no consistent relationship between feelings of restlessness (question 8) and performance on watch.

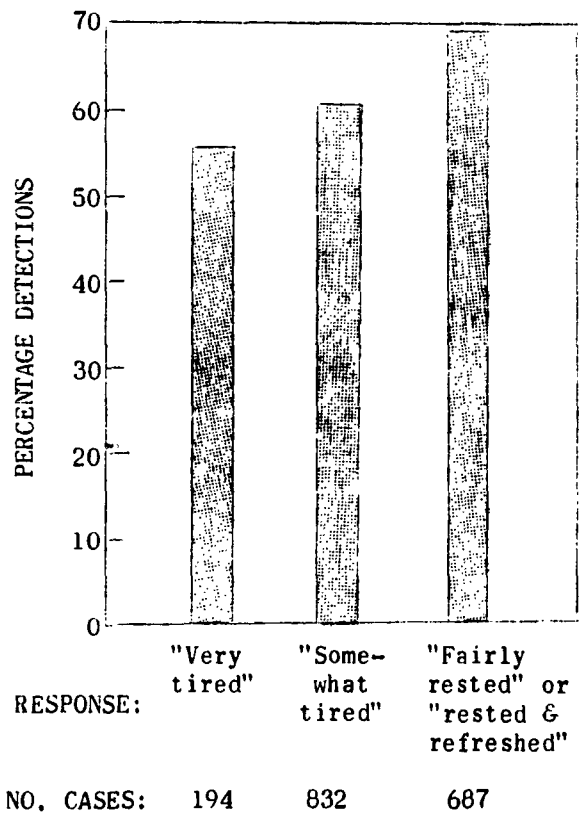

Figure 5. Relationship between response to the question "How do you feel right now?" and detection performance.

Discussion. The results indicated that subjects performed better when they felt rested than when they felt tired. Feelings of tiredness increased systematically during the day. These results are consistent with the findings of the main experiment and of other experiments (Jenkins, 1958) that vigilance performance is poorer on afternoon watches than on morning watches. The results also agree with those of Fraser and Samuel (1956) who found an increase in feelings of "weariness" accompanying a decline in vigilance performance.

The effect of previous sleep on vigilance performance is particularly significant in that the subjects were not experimentally deprived of sleep, but the sleep deprivation occurred naturally. The effects of sleep deprivation on performance are sometimes difficult to demonstrate in laboratory settings because subjects may be inclined to overcompensate for the sleep loss. It seems clear that men who are assigned vigilance tasks should be allowed sufficient off-watch time to obtain adequate sleep.

While the type of work performed prior to coming on watch was unrelated to vigilance performance, the subject's interest in that work may have been important. This finding suggests the possibility that work interests generalize from one job or activity to another. Perhaps men who perform vigilance tasks on watch should be given stimulating activities off watch.

PERSONAL INTERVIEWS

Each subject was privately interviewed by one of the experimenters about a week and a half after the last experimental sesson. The interviews were tape recorded with the knowledge and consent of the subjects. The purposes of the experimenters in conducting these interviews were to obtain information about the subjects' general attitudes toward the vigilance tasks and how these attitudes may have changed from day to day or within a single watch period; to obtain statements of individual preferences for the auditory or visual tasks; to explore generally

the possibilities of uncontrolled factors that might have operated during the experiment unknown to the experimenters; and to determine possible trainable techniques of watchstanding used by the different subjects. The present analysis was concerned only with the first two types of information: general attitude and task preference.

General Attitude and Vigilance Performance. The subjects were classified into three categories on the basis of their attitude toward the tasks as expressed in the interviews:

1. Positive--expressing a cooperative attitude. (N = 18)
2. Neutral --expressing a non-committal attitude. (N = 10)
3. Negative--expressing a dislike for the tasks. (N = 16)

It was not possible to classify ten of the 54 subjects either because they were unresponsive in the interviews and made no definitive statements or because the experimenters could not agree on how they should be classified. The performance scores of the three groups on both the auditory and the visual tasks were compared. The results indicated that the neutral or non-committal group performed poorest on watch (Figure 6). The differences between group means were not great enough to be statistically significant ($p > .05$), but were consistent for the two different tasks.

Task Preference and Vigilance Performance. The subjects were classified into three categories on the basis of their expressed preference for the visual or auditory task:

1. Prefer visual (N = 6)
2. No preference (N = 21)
3. Prefer auditory (N = 21)

Six subjects were unclassifiable. There was a significant relationship ($p < .05$)

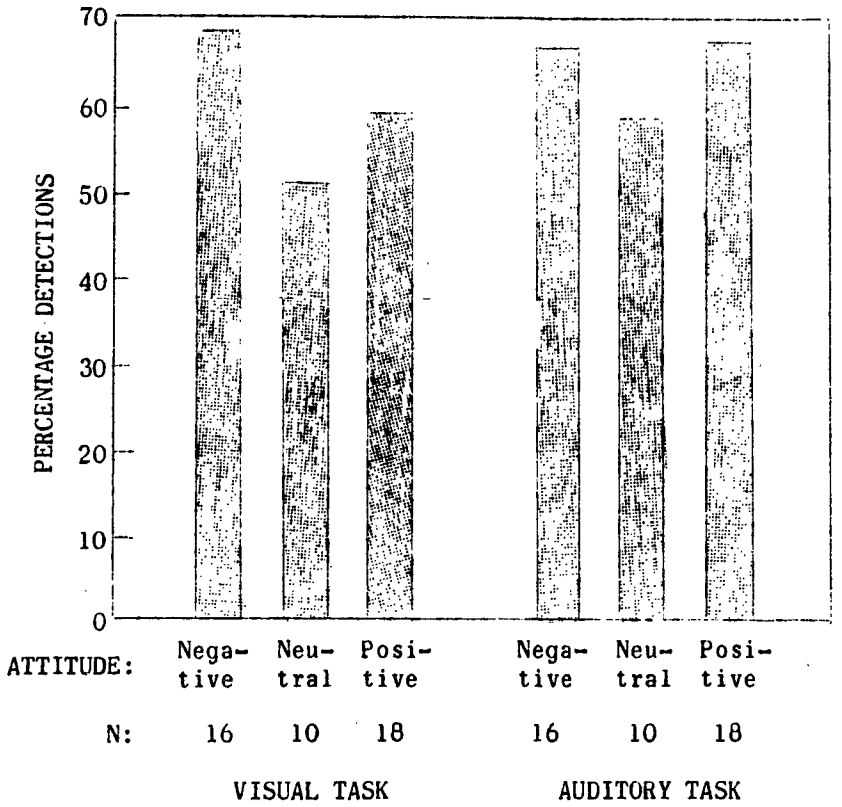

Figure 6. Vigilance performance related to expressed attitude toward task.

between task preference and percentage of signals detected on watch. Subjects performed better on the preferred task than on the non-preferred task (Figure 7). Subjects with no preference performed equally well on both tasks. There was no relationship between task preference and performance under alerted conditions, nor was there any relationship between preference and threshold score.

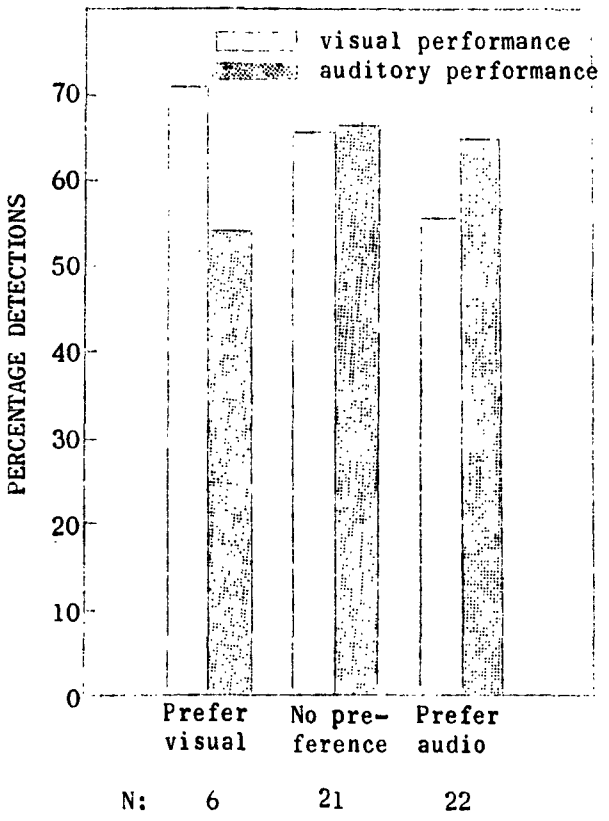

Figure 7. Vigilance performance related to task preference.

Discussion. The effect of general attitude toward the task on vigilance performance was not statistically significant. The effect was consistent, however, for both the auditory and the visual tasks. It was also consistent with the correlation between scores on the MMPI "K" scale and percentage of signals detected on watch. The "K" scale is assumed to measure the degree of guardedness or cautiousness with which persons answer questions about their habits, attitudes, or opinions. The scores on this scale were found to be negatively correlated with vigilance

performance scores. That is, subjects who were non-committal in their responses to the items on the MMPI performed poorer on watch than subjects who responded freely and candidly to those items. It could be predicted, then, that subjects who reported non-committal attitudes toward the vigilance tasks would perform poorer on watch than subjects who candidly reported either positive or negative attitudes.

There was a relationship between task preference and percentage of signals detected on watch such that performance was better on the preferred task. The direction of causality in this relationship is unknown. The subjects may have performed better on one task because they preferred that task to the other, or they may have based their expressions of preference on their estimated performance on one task compared with the other. Evidently, though, subjects did not prefer a task because they were more sensitive to signals on that mode, because there was no relationship between task preference and discrimination thresholds.

REFERENCES

1. Applezweig, M.H. and Moeller, G. The Behavior Interpretation Inventory. New London: Connecticut College, 1958.
2. Bakan, P. Discrimination decrement as a function of time in a prolonged vigil. J. exp. Psychol., 1955, 50, 387-390.
3. Baker, C.H. Attention to visual displays during a vigilance task. II. Maintaining the level of vigilance. Brit. J. Psychol., Feb., 1959.
4. Bowen, H.M. On the appreciation of serial displays. Ph.D. Thesis. University of Cambridge, Cambridge, England, 1956.
5. Buckner, D.N., Harabedian, A., and McGrath, J.J. Human factor problems in anti-submarine warfare, Technical Report 2: A study of individual differences in vigilance performance. Los Angeles: Human Factors Research, Inc., 1960.
6. Fraser, D.C. and Samuel, G.D. Physiological and psychological studies, Part II. Air crew fatigue in long range maritime reconnaissance. Effects on vigilance. Royal Air Force Institute of Aviation Medicine, Report no. 709-10, 1956.
7. Foy, G. A study of the relationships between certain factor analytic ability measures and success in college engineering. Ph.D. Thesis. University of Southern California, Los Angeles, Calif., 1959.
8. Guilford, J.P. Psychometric Methods. New York: McGraw-Hill, 1954.
9. Guilford, J.P., Frick, J.W., Christensen, P.R., and Merrifield, P.R. A factor-analytic study of flexibility in thinking. Reports from the psychological laboratory, University of Southern California, 18, 1957.
10. Guilford, J.P. and Zimmerman, W.S. The Guilford-Zimmerman Temperament Survey--manual of instructions and interpretations. Beverly Hills: Sheridan Supply Co., 1949.
11. Holland, J.G. Human vigilance. Science, 1958, 128, 61-67.

12. Jenkins, H.M. The effect of signal-rate on performance in visual monitoring. Am. J. Psychol., 1958, 71, 647-661.

13. Mackworth, N.H. Researches on the measurement of human performance. Medical Research Council Special Report Series No. 268, 1950.

14. Solandt, D.Y. and Partridge, D.M. Research on auditory problems presented by naval operations. J. Canad. Med. Services, 1946, 3, 323-329.

15. Taylor, J.A. A personality scale of manifest anxiety. J. abnorm. soc. Psychol., 1953, 48, 285-290.

16. Wittenborn, J.R. Factorial equations for tests of attention. Psychometrika, 1943, 8, 19-35.